THUMOS'14 Action Recognition Challenge

http://crcv.ucf.edu/THUMOS14/

Evaluation Setup

UPDATED 8/20/14

The goal of the THUMOS challenge is to recognize and localize a large number of human action classes from open source videos in a realistic setting. The action classes range from daily-life actions (e.g., "Blow Dry Hair" and "Brushing Teeth") to sports actions (e.g., "Driving" and "Golf Swing"). In THUMOS'14, there are two main tasks:

- **Action Recognition**: for a target action class, predict its presence/absence in a given test video associating with a confidence score.
- **Temporal Action Detection**: for a target action class, predict not only its presence in a given video, but also its temporal location (i.e., the starting and ending times of each detected instance).

1 Data

The data resource of THUMOS'14 includes four parts: training data, validation data, background data and test data, which are released in two phases. The training and validation data are released with this document, and test data will be released on 15 August, 2014. The training data is based on the UCF101 [2] action dataset, where videos are temporally trimmed (each video usually contains one instance of the action without irrelevant frames). The remaining three parts (validation, background, and test data) were collected recently and the videos are temporally untrimmed, which makes the tasks more challenging than the THUMOS'13 competition [1].

1.1 Action Recognition Data Set

We provide four datasets for the recognition task [DOWNLOAD LINK]:

- **Training data**: the entire UCF101 action dataset is used for training. It consists of 101 human action categories with **13,320** videos in total. Each category has more than 100 video clips, all of which are temporally trimmed.
- **Validation data**: There are **1,000** videos in total provided for 101 action classes as validation data (each class has exactly 10 videos). In general, there is one primary action class shown in each video; however, some videos may include one or more instances from other action classes. The video-level label information (including the primary and secondary actions for each video) is provided (link) but all videos are temporally untrimmed. The validation set can be used to tune your system or simply as additional training data.

- **Background data**: **2,500** videos, which are verified to make sure they do not include an instance of any of the 101 action classes, are provided as background. Each video is relevant to one of the action classes (even though no background video includes an instance of any of the 101 classes). For instance, the background videos related to the action class "Basketball Dunk" may show the basketball court when the game is not being played. The primary class for each background video is provided (link). The background data may also be used as additional training data.
- **Test data**: **1,574** temporally untrimmed videos are provided as the test data. Some videos may contain one or multiple instances from one or multiple action classes, and some videos may not include any actions from the 101 classes. This set will not be released until August 15, 2014. The ground truth will not be released until after the challenge session at ECCV'14.

1.2 Temporal Action Detection Data Set

For the temporal detection (localization) task, we also provide four datasets. The shared data is mainly the same as the one described in section 1.1, with two major differences: the task is limited to 20 classes (as compared to 101), and the temporal locations of the action instances in 200 validation videos is provided. To further summarized the data:

- **Training data**: a subset of UCF101 action dataset with 20 action classes is used for training. The remaining action classes in UCF101 may also be used for training, if desired. The list of the 20 actions of interest is available here.
- **Validation data**: 200 videos from 20 action classes are provided as the validation set. These videos have the same properties as the validation videos in the recognition task. The validation set could be added to the training data, if desired. The temporal annotations (start and end time) of all instances of the actions occurring in the validation videos is provided (link).
- **Background data**: this is the same background data as the one provided for the recognition task.
- **Test data**: 1574 temporally untrimmed videos are provided as the test data. This data set will not be released until August 15, 2014. The ground truth will not be released until after the challenge session at ECCV'14.

1.3 Pre-Computed Low-level Video Features

The Improved Dense Trajectory Features [3], which are one of the state-of-the-art features for action recognition, are pre-computed and made available for all of datasets (training, validation and background videos). Instead of providing the raw Improved DTF features, the extracted features are quantized using a codebook (with 4000 terms) in order to decrease the size of the shared data (link to download the codebook and features). The spatio-temporal information of each Improved DTF feature is provided.

Each text file in the features folder represents one video, and each row therein denotes one improved DTF feature with the following format:

[Frame_num] [mean_x] [mean_y] [Traj_index] [HOG_index] [HOF_index] [MBH_index]

Frame_num: The frame on which the trajectory ends.
mean_x: The mean value of the x coordinates of the trajectory
mean_y: The mean value of the y coordinates of the trajectory
Traj_index: The codebook term number (0~3999) that the trajectory feature was quantized to.
HOG_index: The codebook term number (0~3999) that the HOG feature was quantized to.
HOF_index: The codebook term number (0~3999) that the HOF feature was quantized to.
MBH_index: The codebook term number (0~3999) that the MBH feature was quantized to.

The same features will be made available for the test videos on August 15. Participants are also encouraged to extract their own features for the contest.

2 Action Recognition Task

2.1 Definition

This task is similar to the conventional action recognition, which is defined as predicting the presence/absence of the action classes in a test video. Specifically, a system shall output a real-valued score indicating the confidence of the predicted presence. As compared to THUMOS'13 [1], the test videos are more challenging since all the videos are temporally untrimmed. In other words, a significant part of a test video may not include any particular action, and multiple instances may occur at different timestamps within the video. In addition, videos that do not contain any of the 101 actions will be a part of the test set.

2.2 Submission Format

Each team is allowed to submit the results of at most five runs. The run with the best performance will be selected as the primary run of the submission and will be used to rank across teams. Each run must be saved in a separate text file with the following format:

```
[test_video_name_1]    [confidence_score_class_1]  [confidence_score_class_2] … [confidence_score_class_m]
[test_video_name_2]    [confidence_score_class_1]  [confidence_score_class_2] … [confidence_score_class_m]
[test_video_name_3]    [confidence_score_class_1]  [confidence_score_class_2] … [confidence_score_class_m]
[test_video_name_4]    [confidence_score_class_1]  [confidence_score_class_2] … [confidence_score_class_m]
.                      .                                .                              .
.                      .                                .                              .
.                      .                                .                              .
[test_video_name_n]    [confidence_score_class_1]  [confidence_score_class_2] … [confidence_score_class_m]
```

Here n = 1,500 (total number of test videos) and m = 101 (total number of test classes). Basically, each row shows the results of one test video, and each column contains the confidence score of presence of the corresponding action class anywhere in the video. The 101 action class labels are available here. The confidence score must be between 0 and 1. A larger confidence value indicates greater confidence to detect the action of interest in a test video. To help with better understanding the format of the submission text file, a sample submission can be seen here. Participants shall strictly follow the submission format, otherwise they incur the risk of not having their results evaluated and posted in the evaluation report and on the challenge website.

2.3 Evaluation Metric

We will use Interpolated Average Precision (AP) as the official measure for evaluating the results on each action class. Given a descending-score-rank of videos for the test class *c*, the *AP*(*c*) is computed as:

$$AP(c) = \frac{\sum_{k=1}^{n}\big(P(k) \times rel(k)\big)}{\sum_{k=1}^{n} rel(k)}, \qquad (1)$$

where *n* is the total number videos, *P*(*k*) is the precision at cut-off *k* of the list, *rel*(*k*) is an indicator function equaling to 1 if the video ranked *k* is a true positive, and to zero otherwise. The denominator is the total number of true positives in the list.

Mean Average Precision (mAP) is the official measure used to evaluate the performance of one run, which is computed as:

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP(c)\,, \qquad (2)$$

where *C* is the total number test classes, which is equal to 101.

3 Temporal Action Detection Task

3.1 Definition

Besides predicting the presence/absence of an action class in a test video, temporal localization is also required in this task. A system should output a real-valued score indicating the confidence of the prediction, as well as the starting and ending time for the given action. A subset of 20 action classes out of 101 will be employed for this task.

3.2 Submission Format

Each team can submit at most 5 runs. The run with the best performance will be selected as the primary run of the submission and will be used to rank across teams. Each run should be saved in a separate text file with the following row format:

[video_name] [starting_time] [ending_time] [class_label] [confidence_score]

Each row has 5 fields representing a single detection. A detector can fire multiple times in a test video (reported using multiple rows in the submission file). The time should be in seconds with one decimal point precision. The confidence score shall be between 0 and 1. A larger confidence value indicates greater confidence in detecting the action of interests in a test video. To help with better understanding the format of the submission text file, a sample submission can be seen here. Participants shall strictly follow the submission format, otherwise they incur the risk of not having their results evaluated and posted in the evaluation report and on the challenge website.

3.3 Evaluation Metric

Interpolated Average Precision (AP) and its mean value (mAP) are the official metrics used to measure the performance of each action class and each run, respectively. A detection is marked as true or false positive based on the time period of overlap with the ground truth time range. The overlap o between the predicted time range R_p and ground truth time range R_{gt} is computed as:

$$o = \frac{R_p \cap R_{gt}}{R_p \cup R_{gt}}. \qquad (3)$$

When o is larger than 50%, the detection is considered correct.

4 Development kit

For the convenience of the participants, a software package composed of the code of the evaluation metrics for the recognition and detection tasks, as well as additional items will be made available (link).

5 Result Submission

Please click here to submit your results to the competition. Each participant may submit the results of up to five runs for each task based on different configurations of their systems. All the results for each task should be zipped into a single file named by [organizationName-taskName.zip]. Within the zipped folder, results from different runs should be placed in separate files named by [Run-Number]. Sample submission files can be seen here.

6 Important Dates

- **Training Data Release 1** (Training+Validation videos, features, evaluation setup): July 18, 2014
- **Training Data Release 2** (Development kit, Temporal Annotations): Week of July 21, 2014
- **Test Data Release:** August 15, 2014
- **Submission Deadline:** August 30, 2014
- **Notebook Paper Due:** September 4th, 2014
- **Challenge Results Notifications:** September 3, 2014
- **Workshop Date:** September 7, 2014
- **Publication Date:** As per conference schedule

References:

[1] Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/ICCV13-Action-Workshop/, 2013.

[2] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. In CRCV-TR-12-01, 2012.

[3] H. Wang and C. Schmid, Action Recognition with Improved Trajectories, in IEEE International Conference on Computer Vision (ICCV), 2013.